PAIN

Pain xx (2002) xxx–xxx

www.elsevier.com/locate/pain

Reproducibility of pain measurement and pain perception

Elisa M. Rosier[a], Michael J. Iadarola[a], Robert C. Coghill[b,*]

[a]*Pain and Neurosensory Mechanisms Branch, National Institute of Dental and Craniofacial Research, National Institutes of Health, Bethesda, MD 20892, USA*

[b]*Department of Neurobiology and Anatomy, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157-1010, USA*

Received 16 October 2001; received in revised form 12 February 2002; accepted 19 February 2002

Abstract

The reproducibility of both the conscious experience of pain and the reproducibility of psychophysical assessments of pain remain critical, yet poorly characterized factors in pain research and treatment. To assess the reproducibility of both the pain experience and two methods of pain assessment, 15 subjects evaluated experimental heat pain during four weekly sessions. In each session, both brief (5 s) and prolonged (90 s) heat stimuli were utilized to determine effects of stimulus duration on reproducibility. Multiple presentations of the brief heat stimuli in each session were used to evaluate effects of response averaging. Both visual analog scales (VAS) and randomized verbal descriptor scales (VDS) were employed to better distinguish variations in the pain experience from variations in pain scale usage. Subjects also rated the intensity of visual stimuli in order to provide an independent assessment of the session-to-session variation in the use of both types of scales. Within-subjects analyses revealed that ratings of visual stimuli exhibited significantly less session-to-session variation than ratings of heat pain. Thus, pain perceptions were more variable than perceptions of visual stimuli after controlling for session-to-session variations in scale usage. Comparisons between scales indicated that intensity ratings acquired with the VAS had significantly smaller session-to-session variation than those acquired with the VDS, although VDS ratings were spread across a larger range of the scale. For both scales, analyses of the effects of stimulus averaging and stimulus duration revealed that averaging multiple assessments of the same stimulus substantially reduces session-to-session variation and that multiple assessments of brief stimuli produce responses which are more reproducible than a single presentation of a prolonged stimulus. However, the VAS was significantly more sensitive to small differences in perceived pain intensity and pain unpleasantness, and did not exhibit some of the order effects present with the VDS. Taken together, these results indicate that the reproducibility of psychophysical ratings of pain can be maximized: (1) by averaging responses to multiple, brief stimuli; (2) by providing subjects with a training period distinct from the study period; and (3) by ensuring that interpretation of scale parameters remains constant over time. Thus, although the experiences of both experimental and clinical pain are highly variable, pain assessment procedures can be structured to minimize session-to-session variability. © 2002 Published by Elsevier Science B.V. on behalf of International Association for the Study of Pain.

Keywords: Pain measurement; Reproducibility; Pain perception; Visual analogue scale; Verbal descriptor scale

1. Introduction

Psychophysical measurement of pain is a critically important aspect of both acute and chronic pain management. In the vast majority of clinical situations, multiple sequential assessments of pain are acquired during the course of treatment. Yet, a given individual's pain rating may vary substantially from measurement to measurement. Such temporal variations can arise from both variation in that individual's *actual pain experience* and variation in how that individual *reports* the experience. Clearly, temporal variations in *measurements* of pain can impede treatment in clinical settings and can reduce statistical power in research settings. Thus, assessment of the temporal variation in pain measurements is a critical component of the validation process of pain scaling procedures. For example, the test–retest repeatability of both visual analog scales (VAS) and verbal descriptor scales (VDS) have been examined in detail with correlational methods. Both scales have been shown to exhibit very high test–retest correlations (Gracely et al., 1978; Price et al., 1983). However, the use of such correlational methods to assess test–retest repeatability has been criticized on the grounds that the correlation between two measurements only assesses the strength of the relationship between these measures, rather than the actual agreement between them (Altman and Bland, 1983; Bland and Altman, 1986). An alternative statistic, the coefficient of repeatability, has instead been proposed to assess the agree-

* Corresponding author. Tel.: +1-336-716-4284; fax: +1-336-716-4534.
E-mail address: rcoghill@wfubmc.edu (R.C. Coghill).

0304-3959/02/$20.00
PII: S0304-3959(02)00048-9

PAIN 4704
NA 4289

ment between two separate measures of a given phenomenon (British Standards Institution, 1979; Altman and Bland, 1983; Bland and Altman, 1986).

In an effort to specifically assess the repeatability of VAS measurements of pain using non-correlational techniques, Yarnitsky et al. (1996) examined a large group ($n = \sim 30$) of normal volunteers across four weekly measurement sessions. Each subject evaluated three levels of heat pain, with each level being defined by an offset of 1.5, 3.0, or 4.5°C above that subject's pain threshold. Coefficients of repeatability for these suprathreshold pain ratings were very large (3.8–4.7) relative to the range of the scale (10). In the case of a coefficient of repeatability of 3.8, there would be a 95% probability that a pain experience that was rated as 5/10 in one week could be rated anywhere between 1.2/10 and 8.8/10 in the following week, even though the pain experience did not change from one week to the next. Accordingly, Yarnitsky et al. (1996) concluded that the VAS has a poor repeatability and questioned the use of the VAS for evaluation of pain across multiple sessions. However, several aspects of these conclusions are problematic (Price, 1997). First, the perception of the experimental stimulus may have been highly variable, despite the fact that the stimulus itself was quite consistent. Yarnitsky et al. (1996) only employed a single pain assessment tool, and had no means of independently assessing the week-to-week variation of pain *perceptions*. Therefore, they could not distinguish whether variations in VAS ratings were attributable to measurement errors inherent in the VAS or could be attributed to variations in the percept. Second, they made no attempt to assess the repeatability of the VAS against other more perceptually stable stimuli. Such a procedure could have provided further insight as to whether variations in the percept or variations in the measure accounted for the majority of session-to-session differences. Third, in linking the suprathreshold stimulus intensities to pain threshold, they clearly exacerbated session-to-session variation in the ratings of suprathreshold stimuli by convolving them with the known variability of the pain threshold (Yarnitsky et al., 1995). Finally, they provided no insight as to strategies for minimizing session-to-session variation.

The objectives of the present study are to better characterize potential sources of session-to-session variation in ratings of pain and to develop strategies to minimize these variations. Two separate approaches were taken to distinguish session-to-session variations in pain *perceptions* from session-to-session variations in *measurements* of those perceptions. First, a randomized VDS was employed in addition to the VAS (Gracely et al., 1978; Heft et al., 1980). Such randomized category scales require ratings to be based solely on the semantic content of a descriptor rather than its spatial location in a list of descriptors. Accordingly, the use of such an independent assessment of pain can provide insight as to the magnitude of session-to-session *perceptual* variation. Secondly, perceptually stable visual stimuli were assessed with both VAS and VDS in order to provide an assessment of the degree to which variations in scale usage could contribute to session-to-session variation. To develop strategies to optimize repeatability, session-to-session variations in pain measures from each scale were compared to determine if one type of scale yielded more reproducible results. Similarly, different methods of presenting heat stimuli were employed to gain insights as to how session-to-session variation could be minimized by repeated stimulus presentations/session or by increasing stimulus duration.

2. Methods

2.1. Subjects

Eight male and seven female volunteers participated in this study following recruitment from the NIH normal volunteer program. Subjects ranged in age from 21 to 36 years. Eleven subjects were white (four females, seven males) and four were black (three females, one male). All subjects were native English speakers, right handed, and healthy, without history of pain or neurological disorders. No subject had previously participated in a psychophysical study of pain using either the VAS or VDS. Subjects gave informed consent acknowledging that they understood: (1) that the experiment involved presentation of heat-induced pain, (2) that the methods to be used were clearly explained and understood, (3) that no tissue damage would result from stimulation, and (4) that they were free to terminate stimulation or to withdraw from the study at any time. All procedures were approved by the Institutional Review Board of the National Institute of Dental Research.

2.2. Schedule of data acquisition

Subjects participated in four sessions, each separated by exactly 7 days. In all sessions, subjects rated brief heat stimuli, prolonged heat stimuli, visual stimuli, and provided an evaluation of the worst physical pain that they had previously experienced (Table 1).

2.3. Sensory stimuli

2.3.1. Heat stimuli

Graded heat was selected as the primary noxious stimulus for this investigation since it has been widely used in experimental studies of pain and since it can be readily delivered in a highly accurate fashion. All stimuli were delivered at a 6°C/s rise rate with a 1-cm diameter, resistance heated, feedback-controlled stimulator and were presented in random order on the subject's non-dominant ventral forearm. In order to be consistent over sessions and between subjects, stimuli were delivered to eight sites defined by a 4×2 grid of eight ink marks centered at the measured midpoint of each subject's forearm.

Two types of heat stimuli (brief and prolonged) were

Table 1
Session schedule[a]

Experimental task	Scale used	
	Visual analogue scale	Verbal descriptor scale
Worst pain rating	1 rating/session	1 rating/session
Training	(35°C, 43–49°C)/session	(35°C, 43–49°C)/session
Brief (5 s) Heat	3 × (35°C, 43–49°C)/session	3 × (35°C, 43–49°C)/session
Prolonged (90 s) Heat	1 × (35, 45, 47, and 49°C)/session	1 × (35, 45, 47, and 49°C)/session
Visual	2 × (8 shades of grey)/session	2 × (8 shades of grey)/session

[a] Subjects participated in all five tasks in every session. In the case of the worst pain and the training tasks, only one scale/session was used, whereas both scales were used in each session for all other tasks.

employed to assess the effects of stimulus duration on the reproducibility of pain. The brief heat stimuli were 5 s in duration while the prolonged heat stimuli were 90 s in duration. Each brief heat stimulus was applied to a single marked skin region, while each prolonged stimulus was produced by sequential placement of the stimulator to each of the eight marked skin regions (5 s/region, 0.5 s interval). The sequential placement of the stimulator during the prolonged stimuli served to minimize confounds due to sensitization or habituation. The brief heat stimuli (35, 43, 44, 45, 46, 47, 48, and 49°C) were each applied six times/session in a random order. The prolonged heat stimuli (35, 45, 47, and 49°C) were each applied two times/session in a random order.

To minimize anxiety and to familiarize subjects with the heat stimuli, subjects underwent a brief training block at the beginning of every session. During this block, subjects applied each of the eight brief heat stimuli to their own dominant forearm in a fixed, ascending order. Responses to these stimuli were not examined further.

2.3.2. Visual stimuli

In order to better distinguish session-to-session variations in scale usage from session-to-session variations in perception, subjects rated visual stimuli in addition to heat stimuli. The visual stimuli were chosen to provide a highly reproducible non-painful sensory experience. These stimuli consisted of eight shades of gray (0, 14, 29, 44, 68, 72, 86, and 100% black; Fig. 1), which were presented four times/session in a random order. Each stimulus occupied a 1 cm × 1 cm shaded area on a 8.5 × 11 inch sheet of white paper (one stimulus/sheet).

2.3.3. Recollection of worst pain

During the first session, subjects were asked to remember the worst physical pain of their life and rate its intensity and unpleasantness. During each subsequent session, the subject was reminded of his/her worst pain and asked to rate it again.

2.4. Psychophysical assessment

Both VAS and VDS were used for psychophysical assessment of the thermal and visual stimuli. These two distinct scales were employed in order to better differentiate session-to-session variations in *perceptions* from session-to-session variations in *measurements* of those perceptions.

The VAS consisted of a 15-cm sliding scale device (Parisian Novelty Company, 3510 S. Western Avenue, Chicago, IL 60609, USA) which has been described in detail and validated previously (Price et al., 1994). Movement of the slider exposes a uniformly red bar to the subject and a numerical scale (range 0–10) to the investigator. Subjects were instructed to separately rate pain intensity and pain unpleasantness with this device exactly as described in Price et al. (1989).

The VDS consisted of previously validated word lists used for pain intensity and pain unpleasantness assessment (Gracely et al., 1978; Heft et al., 1980; Table 2). At the beginning of each session, subjects ranked individual descriptors (on index cards) by sorting them in ascending order of intensity and unpleasantness. However, for each VDS rating, subjects picked a descriptor from a *randomized* list. Accordingly, no spatial cues were provided during descriptor presentation; thus, word meaning alone provided the sole indication of descriptor magnitude.

For statistical analysis of VDS ratings, the rank order from the first session alone was used to later convert word choices from all four sessions to numerical values. This

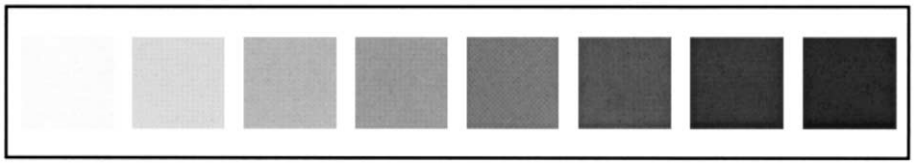

Fig. 1. Shades of gray. Shaded squares (1 cm × 1 cm) were used for visual stimuli. During the rating tasks, each shade was presented separately on a standard sheet of paper.

Table 2
Verbal descriptors for pain intensity, pain unpleasantness, and visual stimuli[a]

Intensity	Unpleasantness	Visual
Barely strong	Annoying	Almost black
Clear-cut	Distressing	Almost gray
Extremely intense	Intolerable	Almost white
Extremely weak	Miserable	Average gray
Faint	Slightly annoying	Barely black
Intense	Slightly distressing	Barely gray
Mild	Slightly intolerable	Barely white
Moderate	Slightly miserable	Black
Slightly intense	Slightly unpleasant	Extremely gray
Slightly moderate	Unpleasant	Faintly gray
Strong	Very annoying	Mildly gray
Very intense	Very distressing	Slightly black
Very mild	Very intolerable	Slightly gray
Very weak	Very miserable	Very gray
Weak	Very unpleasant	White

[a] At the beginning of each session, subjects ranked descriptors in order of increasing magnitude, but picked a descriptor from a randomized list during each rating. Note that descriptors are currently presented in alphabetical order.

process was done on a subject-by-subject basis, so the ranks reflect each individual's interpretation of the descriptors. In order to facilitate comparisons with VAS ratings, VDS ranks were multiplied by 0.667 in order to produce responses with the same range (0–10) as the VAS. This simple ranking procedure yields an ordinal scale based on semantic content and provides a measurement strategy as distinct as possible from the ratio-scaled VAS.

VAS and VDS ratings of brief and prolonged heat stimuli were obtained in a completely randomized order. Furthermore, ratings of intensity and unpleasantness were also obtained in a randomized order. However, for a given stimulus, both intensity and unpleasantness ratings were obtained with the same scale.

Ratings of visual stimuli were also obtained with both VAS and VDS. In the case of the VAS, subjects were instructed that the lower extreme of the scale indicated a complete absence of shading, while the upper extreme of the scale indicated that the shaded area was completely black, and that positions in between were proportional to the darkness of the gray region. For VDS ratings of gray intensity, a set of descriptors analogous to those used for pain intensity ratings was employed (Table 2). As in the case of pain ratings, VAS and VDS ratings of the visual stimuli were obtained in a completely randomized order.

2.5. Statistical analysis

VAS and VDS responses were examined identically in all statistical analyses. To quantitatively characterize session-to-session variation in VAS and VDS responses to a given stimulus, the session-to-session difference was calculated as the absolute value of the difference of the responses between adjacent weeks (i.e. session-to-session difference = /week 1 rating − week 2 rating/). The session-to-session difference was selected instead of the coefficient of repeatability (as defined by the British Standards Institute, 1979) in order to provide a more straight-forward and interpretable measure of how much session-to-session variation existed within a given stimulation/measurement paradigm. To produce a single index reflecting all of the session-to-session variation, the session-to-session differences were then averaged across the three pairs of adjacent weeks (i.e. week1–week2, week2–week3, week3–week4), and averaged across all temperatures >35°C or gray levels >0% black. For within modality analyses of brief heat, prolonged heat, and visual stimuli, these averaged session-to-session differences were then examined with a single-factor, within-subjects analysis of variance (ANOVA) to determine effects attributable to scale type. For comparisons across different stimulus paradigms, two factor, within-subjects ANOVAs were performed to identify changes in session-to-session differences dependent upon stimulus paradigm and/or scale type.

Since session-to-session difference scores provide no information as to the presence of systematic changes in ratings over time (i.e. habituation or sensitization) or about the sensitivity of each scale, another series of analyses were performed separately on both VAS and VDS data. Two factor, within-subjects ANOVAs were used to identify systematic changes in ratings over time, and to assess sensitivity to different magnitudes of stimuli. Univariate contrast analyses were then performed between adjacent pairs of stimulus intensities (i.e. 35 vs. 43°C, 43 vs. 44°C, etc) to determine if small differences in stimulus intensity were successfully detected. Similar analyses were performed between adjacent weeks to identify when systematic changes in ratings occurred.

3. Results

3.1. Reproducibility of heat pain ratings

The variability of heat pain ratings across different experimental sessions is summarized in Fig. 2. The degree of this variation, expressed as the average of the absolute value of session-to-session differences in pain ratings, is substantial, but differs according to the type of scale used to assess pain and stimulus paradigm used to elicit pain.

When subjects used the VAS to evaluate brief heat stimuli, pain intensity ratings exhibited significantly smaller session-to-session variation than those obtained by VDS (INT, $F(1,15) = 8.07$, $P < 0.0124$), while session-to-session variation of VAS ratings of unpleasantness was indistinguishable from those obtained with the VDS (UNP, $F(1,15) = 1.41$, $P < 0.2541$, Fig. 2A). In the case of the VAS ratings, these session-to-session variations in individual scores did not occur in a systematic fashion indicative of habituation or sensitization (Fig. 3). In other

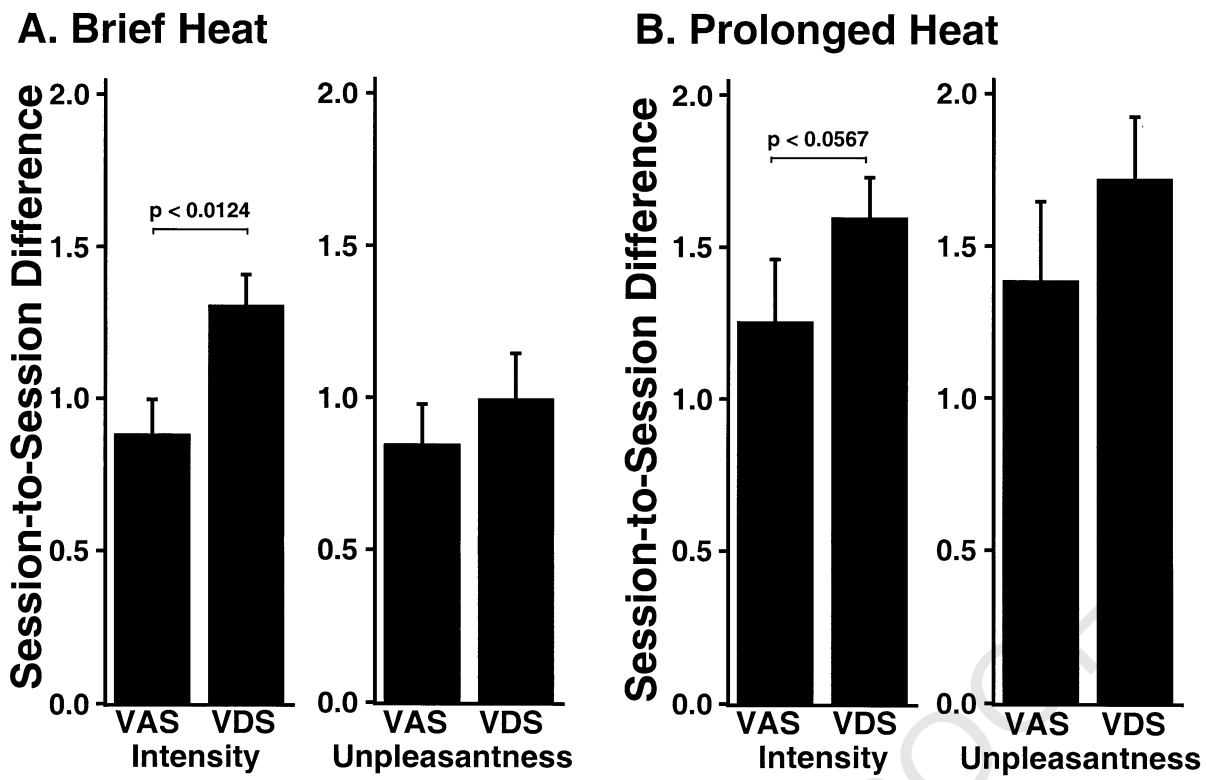

Fig. 2. Session-to-session differences in heat pain. VAS ratings of the intensity of brief heat pain stimuli exhibited significantly less session-to-session variability than those obtained by VDS. In the case of ratings of the prolonged heat stimuli, a similar trend was evident.

words, some individuals' ratings of a given temperature went up from one session to the next, while other individuals' ratings went down. Thus, when averaging ratings across the entire group of subjects, the net session-to-session change is not distinguishable from 0, and there are no statistically significant changes in ratings across sessions (INT, $F(3,42) = 0.97$, $P < 0.4149$; UNP, $F(3,42) = 1.41$, $P < 0.2529$). In contrast, VDS ratings of brief heat stimuli exhibited significant, systematic decreases across sessions (INT, $F(3,42) = 90.9$, $P < 0.0001$; UNP $F(3,42) = 29.59$, $P < 0.0001$, Fig. 3).

Ratings of prolonged heat stimuli also exhibited substantial session-to-session variation (Fig. 2B). In contrast to the ratings of brief heat stimuli, both VAS and VDS ratings of pain intensity and pain unpleasantness exhibited statistically indistinguishable session-to-session differences, although VAS ratings of intensity exhibited a tendency to be less variable than VDS ratings (INT, $F(1,15) = 4.26$, $P < 0.0567$; UNP, $F(1,15) = 1.50$, $P < 0.2393$, Fig. 2B). Neither VAS nor VDS intensity or unpleasantness ratings systematically changed across sessions (VAS INT, $F(3,36) = 1.65$, $P < 0.1946$; VAS UNP, $F(3,36) = 2.35$, $P < 0.0888$; VDS INT, $F(3,36) = 1.30$, $P < 0.2901$; VDS UNP, $F(3,36) = 1.61$, $P < 0.2040$; Fig. 4).

3.2. Sensitivity of VAS and VDS measurements of heat pain

Statistically significant effects of stimulus temperature on perceived pain intensity and pain unpleasantness were readily detected in both VAS and VDS measurements of brief heat stimuli (Fig. 3; Table 3). However, only the VAS detected differences in perceived intensity and unpleasantness between all adjacent pairs of stimulus temperatures (Table 3). In contrast, the VDS only detected differences between 43 and 44° and 46 and 47°C for intensity ratings and differences between 46 and 47°C and 47 and 48°C for unpleasantness ratings (Table 3).

For prolonged heat stimuli, statistically significant effects of stimulus temperature on perceived pain intensity and pain unpleasantness were detected by both VAS (INT, $F(3,36) = 53.15$, $P < 0.0001$; UNP, $F(3,36) = 42.46$, $P < 0.0001$) and VDS (INT, $F(3,36) = 34.92$, $P < 0.0001$; UNP, $F(3,36) = 18.16$, $P < 0.0001$; Fig. 4). Both scales detected differences between all adjacent pairs of stimuli (i.e. 35–45°, 45–47°, and 47–49°C, all comparisons significant at $P < 0.0132$, Fig. 4).

3.3. Effects of stimulus averaging and stimulus duration on the reproducibility of heat pain ratings

Presenting each brief heat stimulus multiple times within each session significantly reduced session-to-session variation, regardless of the scale used to assess pain intensity (VAS $F(3,45) = 4.87$, $P < 0.0051$; VDS $F(3,45) = 4.82$, $P < 0.0054$; Fig. 5). For both scales, an average of three presentations was significantly less variable than an average

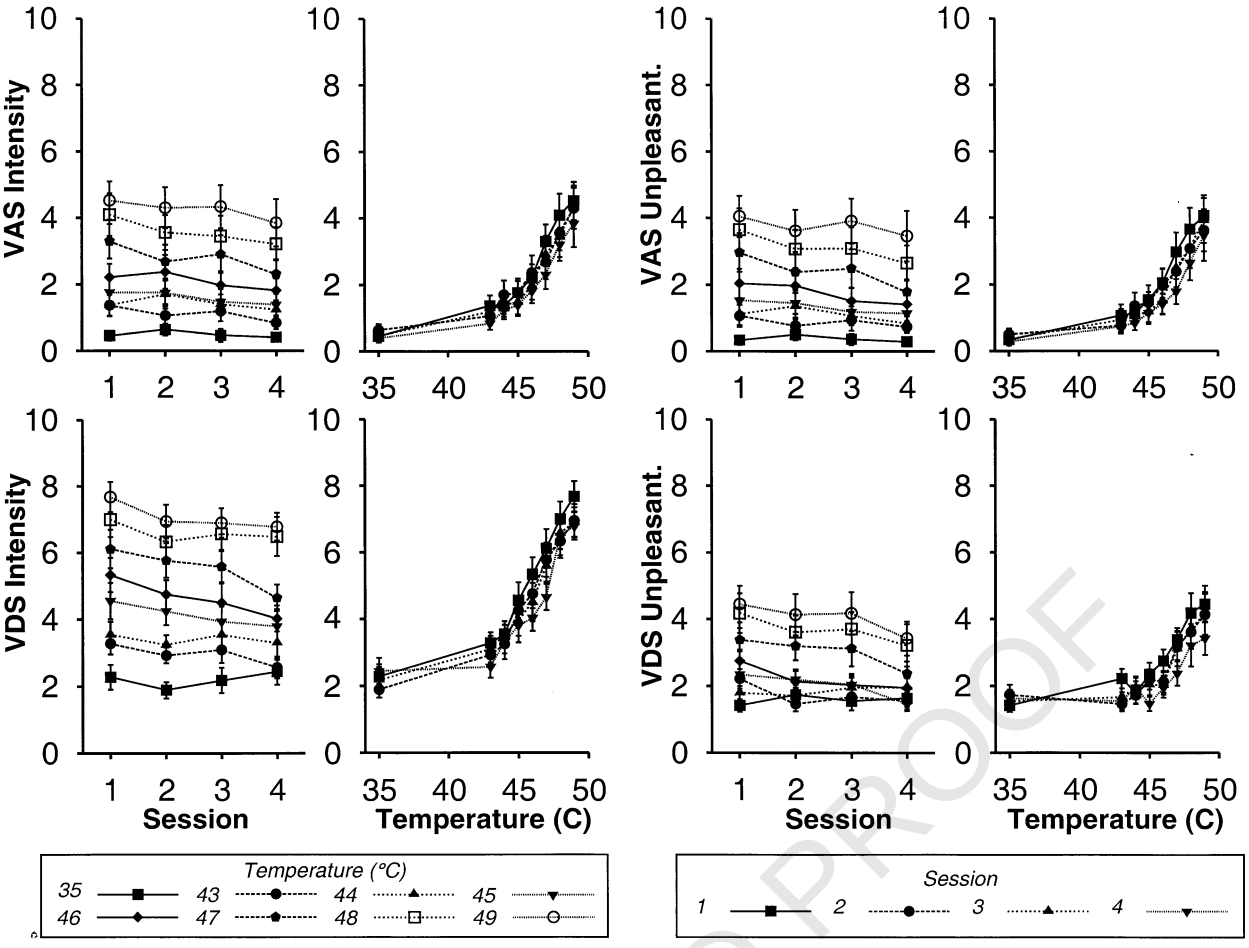

Fig. 3. Brief (5 s) heat ratings across testing sessions. Ratings are presented by session (left column) and by temperature (right column). Ratings obtained with VDS exhibited slight, but significant decreases over time. In contrast, VAS ratings did not change in a statistically reliable fashion over time. Subjects were able to evaluate pain intensity and pain unpleasantness well with both scales, although ratings obtained with the VAS were more sensitive to smaller differences in stimulus intensities.

of two presentations (VAS $F(1,15) = 11.91$, $P < 0.0036$; VDS $F(1,15) = 12.79$, $P < 0.0028$), while an average of two presentations exhibited significantly less session-to-session variation than a single presentation (VAS $F(1,15) = 16.74$, $P < 0.0001$; VDS $F(1,15) = 8.76$, $P < 0.0097$). In the case of the VAS, but not the VDS, an average of three presentations also was significantly less variable in terms of session-to-session variation than one presentation of the prolonged stimulus (VAS $F(1,15) = 4.99$, $P < 0.0411$; VDS $F(1,15) = 2.82$, $P < 0.1138$; Fig. 5)

3.4. Visual stimuli

Session-to-session variations in VAS ratings of visual stimuli were statistically indistinguishable from those of VDS ratings ($F(1,15) = 2.94$, $P < 0.1072$; Fig. 6). Both VAS and VDS detected significant changes in the perceived gray intensity of the visual stimuli (VAS: $F(7,98) = 300.21$, $P < 0.0001$; VDS: $F(7,98) = 237.69$, $P < 0.0001$; Fig. 6). Differences between all shades of gray were detected by both the VAS and the VDS (all comparisons significant at $P < 0.0001$). Neither the VAS nor the VDS ratings exhibited a systematic variation across sessions (VAS: $F(3,42) = 0.75$, $P < 0.5276$; VDS: $F(3,42) = 1.86$, $P < 0.1506$).

3.5. Differences in the reproducibility of visual and heat stimuli

Across all types of stimuli, intensity ratings obtained by the VAS exhibited significantly smaller session-to-session variation than ratings obtained by the VDS ($F(1,15) = 13.31$, $P < 0.0024$; Fig. 7). Regardless of the scale used, ratings of visual stimuli exhibited significantly less session-to-session variation than ratings of both long heat stimuli (INT, $F(1,15) = 24.96$, $P < 0.0002$) and brief duration heat stimuli ($F(1,15) = 31.88$, $P < 0.0001$).

3.6. Reproducibility of recollections of worst pain

Each subject rated their worst physical pain once every session, alternating between the use of the VAS and the

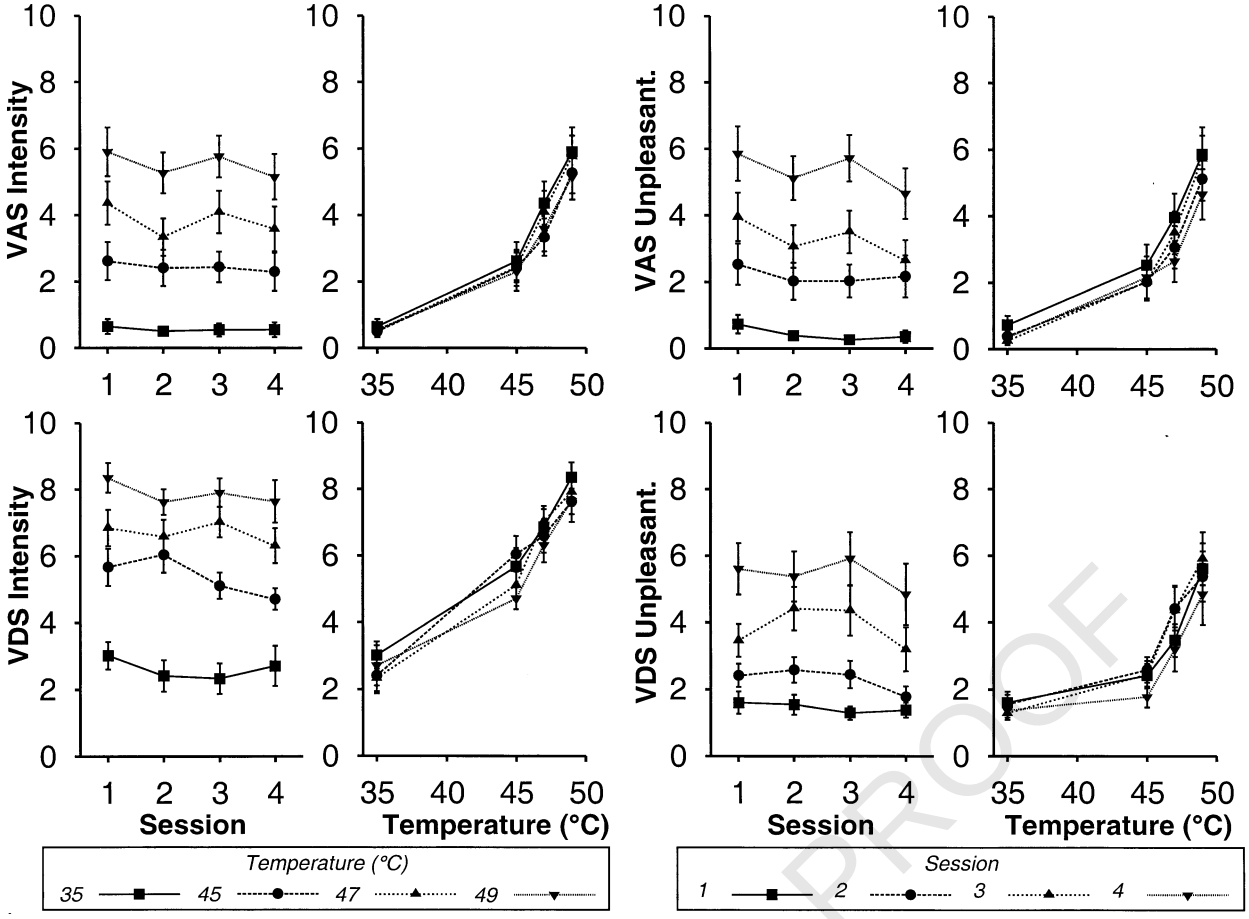

Fig. 4. Prolonged (90 s) heat ratings across testing sessions. Ratings are presented by session (left column) and by temperature (right column). In contrast to ratings of brief heat stimuli, neither VAS nor VDS ratings exhibited systematic changes over time, and subjects were able to detect all differences among stimulus intensities.

Table 3
Effects of stimulus temperature and session on psychophysical ratings of brief heat stimuli (statistically significant effects are presented in bold)

	VAS		VDS	
	Intensity	Unpleasantness	Intensity	Unpleasantness
Session	$F = 0.97$	$F = 1.41$	$F = 90.90$	$F = 29.59$
	$P < 0.4149$	$P < 0.2529$	$P < 0.0001$	$P < 0.0001$
Temperature	$F = 37.76$	$F = 28.97$	$F = 6.04$	$F = 3.97$
	$P < 0.0001$	$P < 0.0001$	$P < 0.0001$	$P < 0.0007$
35 vs. 43	$F = 47.49$	$F = 36.22$	$F = 0.09$	$F = 0.96$
	$P < 0.0001$	$P < 0.0001$	$P < 0.7650$	$P < 0.3449$
43 vs. 44	$F = 40.81$	$F = 33.72$	$F = 5.04$	$F = 0.31$
	$P < 0.0001$	$P < 0.0001$	$P < 0.0415$	$P < 0.5868$
44 vs. 45	$F = 52.78$	$F = 38.50$	$F = 1.50$	$F = 0.20$
	$P < 0.0001$	$P < 0.0001$	$P < 0.2407$	$P < 0.6630$
45 vs. 46	$F = 41.01$	$F = 35.81$	$F = 2.02$	$F = 0.01$
	$P < 0.0001$	$P < 0.0001$	$P < 0.1769$	$P < 0.9096$
46 vs. 47	$F = 50.69$	$F = 37.67$	$F = 10.12$	$F = 12.34$
	$P < 0.0001$	$P < 0.0001$	$P < 0.0067$	$P < 0.0034$
47 vs. 48	$F = 40.18$	$F = 30.36$	$F = 3.27$	$F = 7.72$
	$P < 0.0001$	$P < 0.0001$	$P < 0.0920$	$P < 0.0148$
48 vs. 49	$F = 25.06$	$F = 22.38$	$F = 2.25$	$F = 1.75$
	$P < 0.0002$	$P < 0.0003$	$P < 0.1560$	$P < 0.2073$

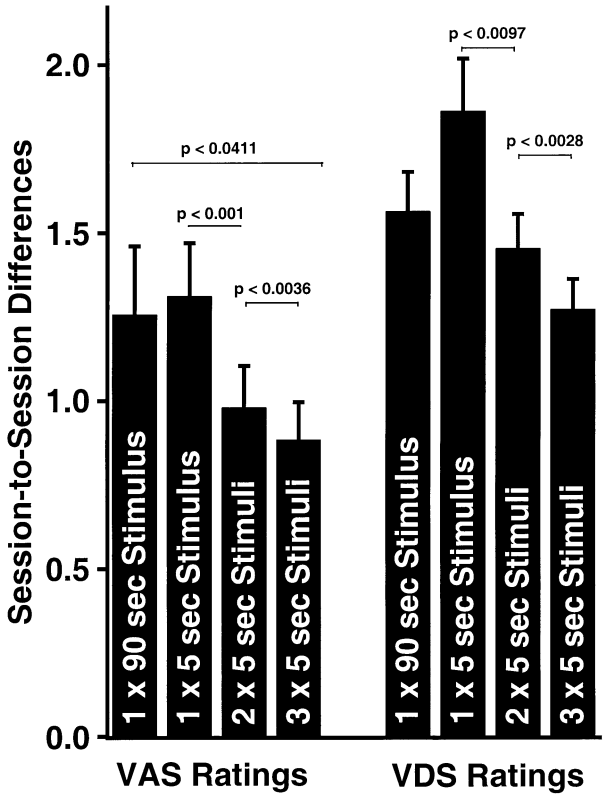

Fig. 5. Effects of averaging on the reproducibility of heat stimuli. Regardless of the scale used for ratings, multiple presentations and assessments significantly reduced session-to-session differences in pain ratings.

VDS each session. Neither VAS nor VDS detected systematic changes in worst pain intensity or unpleasantness ratings across a 2-week interval (VAS: INT, $F(1,14) = 2.32$, $P < 0.1501$; UNP, $F(1,14) = 0.01$, $P < 0.9423$; VDS: INT, $F(1,12) = 0.81$, $P < 0.3870$; UNP, $F(1,12) = 2.87$, $P < 0.1160$; Fig. 8). Both scales exhibited statistically indistinguishable session-to-session variation in both intensity ($F(1,12) = 0.53$, $P < 0.4809$) and unpleasantness ratings ($F(1,12) = 0.63$, $P < 0.4425$, Fig. 8).

3.7. Miscellaneous aspects of the verbal descriptor scale

Subjects' interpretation of the relative rank order of the 15 intensity and the 15 unpleasantness words changed substantially across sessions. For the intensity descriptors, the mean number of discrepancies in rankings between sessions 1 and 2 was 5.93 ± 1.21, between sessions 2 and 3 was 4.06 ± 0.90, and between sessions 3 and 4 was 2.60 ± 0.87 (note that the smallest number of discrepancies would be 2). Accordingly, the consistency of intensity rankings improved significantly over sessions ($F(2,28) = 5.95$, $P < 0.0070$). For unpleasantness rankings, the mean number of discrepancies between sessions 1 and 2 was 5.80 ± 1.03, between sessions 2 and 3 was 3.86 ± 1.07, and between sessions 3 and 4 was 4.26 ± 1.20. There was no significant improvement in the consistency of unpleasantness descriptor ranking across sessions ($F(2,28) = 1.69$, $P < 0.2019$).

Analysis of word choice frequency for VDS ratings of brief heat stimuli showed that, on average, subjects used about 9.17 ± 0.23 intensity words and 6.24 ± 0.37 unpleasantness words in a single session. Five intensity descriptors (faint, intense, mild, moderate, and weak) were used as ratings in 55% of the responses. In contrast, only four unpleasantness descriptors (annoying, slightly annoying, slightly unpleasant, and unpleasant) accounted for 78% of the responses.

4. Discussion

Ratings of pain intensity and pain unpleasantness vary markedly over time within individual subjects, even when obtained within the confines of a carefully controlled experimental situation. This substantial temporal variation can be divided into two general categories. First, the actual experience of pain can vary over time despite the fact that the physical stimulus remains constant. Second, the manner in which subjects report their pain or use the pain scales can vary from one session to the next. The determination of the relative contribution of each of these general categories of temporal variation has critical implications for both experimental design and clinical assessment of pain.

4.1. Test–retest correlations vs. session-to-session differences

The quantification and evaluation of temporal variation of pain ratings has historically been accomplished via the use of test–retest correlations. Both VDS and VAS scales have been demonstrated to have very high test–retest correlations ($r = 0.99$ VDS; $r = 0.97$ VAS) when carefully controlled experimental stimuli are employed (Gracely et al., 1978; Price et al., 1983). The use of such test–retest correlations for the evaluation of reproducibility has been criticized, however, for potential insensitivity to shifts in response slope or response offset which may occur over time (British Standards Institution, 1979; Altman and Bland, 1983; Bland and Altman, 1986). Our findings underscore this weakness of the test–retest correlation procedure. For example, both VAS and VDS measures of brief heat pain had highly significant test–retest correlation coefficients (VAS: $r = 0.84$, $P < 0.0001$; VDS: $r = 0.76$, $P < 0.0001$). In contrast, the average session-to-session differences of these ratings were approximately 20.7% of VAS and 18.5% of VDS ratings of 49°C stimuli, respectively.

To circumvent the problems inherent in test–retest correlations, a confidence interval-based criterion has been proposed as a measure of reproducibility, and has been employed in a previous study of the reproducibility of pain (Yarnitsky et al., 1996). Although this criterion provides a reasonable measure of the reproducibility of a stimulus, it provides no direct descriptive information on the

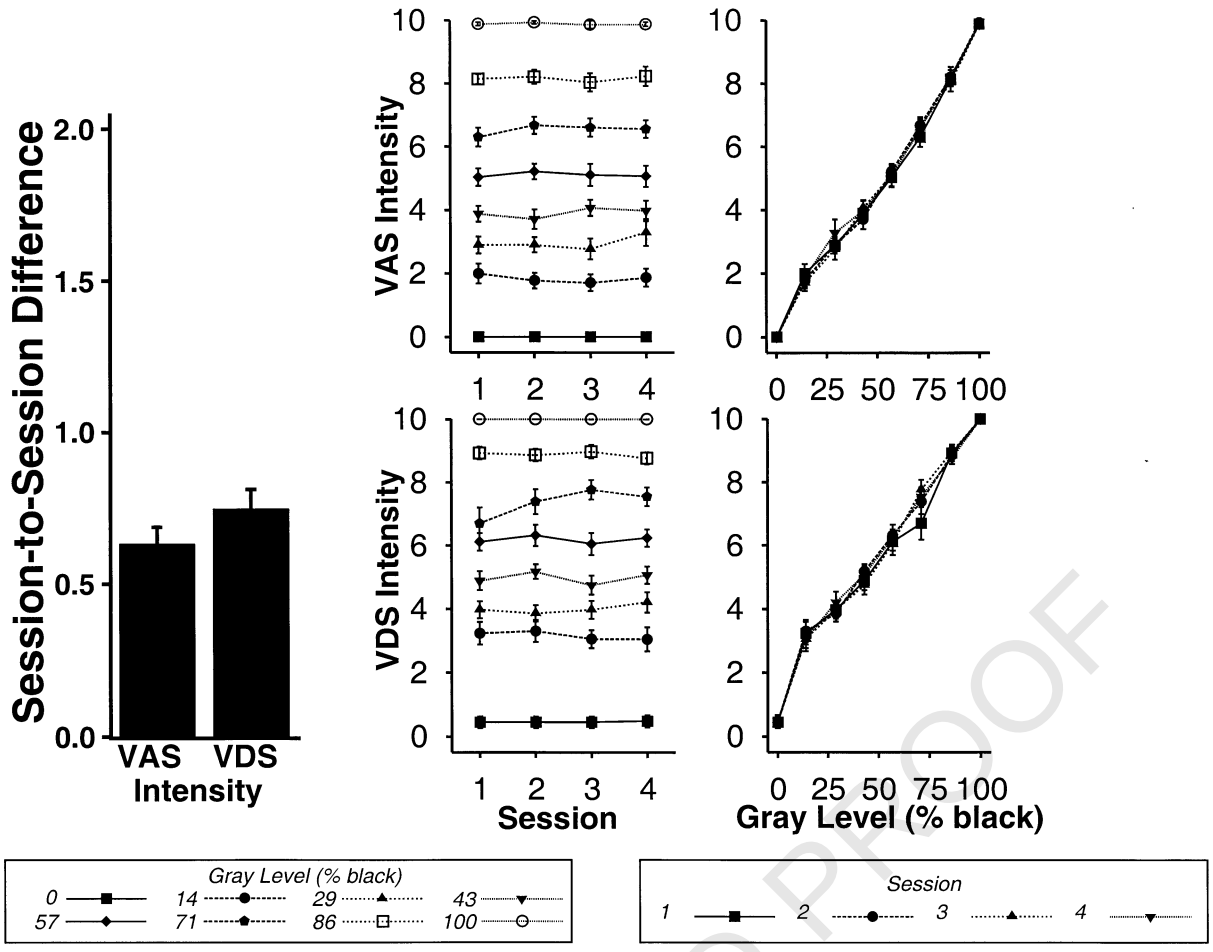

Fig. 6. Session-to-session differences and intensity ratings of visual stimuli. Ratings obtained by both scales were sensitive to small differences in stimulus intensity and had statistically indistinguishable session-to-session differences.

magnitude of session-to-session differences. Accordingly, we have chosen to describe the temporal variation in pain ratings with the relatively simple metric of the session-to-session difference to facilitate the generalizibility of these findings.

4.2. Temporal variations in the pain experience

A substantial portion of the session-to-session variations in ratings of pain intensity and unpleasantness appears to arise from temporal variations in the actual pain experience. Session-to-session variations in ratings of the visual stimuli were significantly smaller than session-to-session variations in ratings of any of the heat pain stimuli, regardless of the scales used (Fig. 7). If temporal variations in scale usage were constant across both stimulus modalities, then a large portion of the difference between session-to-session variations in heat pain ratings and session-to-session variations in ratings of visual stimuli is due to variations in the actual experience of pain. If one next makes the extreme assumption that 100% of the session-to-session variation in the ratings of visual stimuli is due to variation in scale usage, the percentage of variation in brief heat pain ratings attributable to session-to-session differences in the pain experience can be approximated by simply subtracting the session-to-session difference in visual ratings from that of brief heat pain ratings (after normalizing to account for the differing ranges of each scale). Thus, a minimum of approximately 72% of the session-to-session variations in the VAS and 65% of those in the VDS ratings of brief heat stimuli could be attributed to variations in the actual pain experience. (Note that both session-to-session differences are derived from two repetitions of each stimulus). Since this is an overestimate of the perceptual stability of the visual stimuli, larger percentages of session-to-session variations in heat pain ratings are likely due to variations in the actual pain experience.

In the present investigation, temporal variations in the actual experience of pain evoked by the heat stimulus can arise from a variety of physical, physiological, and psychological variables. Physical variables, however, were held as constant as possible. For example, it is highly unlikely that variations in the temperature of the heat stimuli can account for a large portion of the variability in the pain experience as

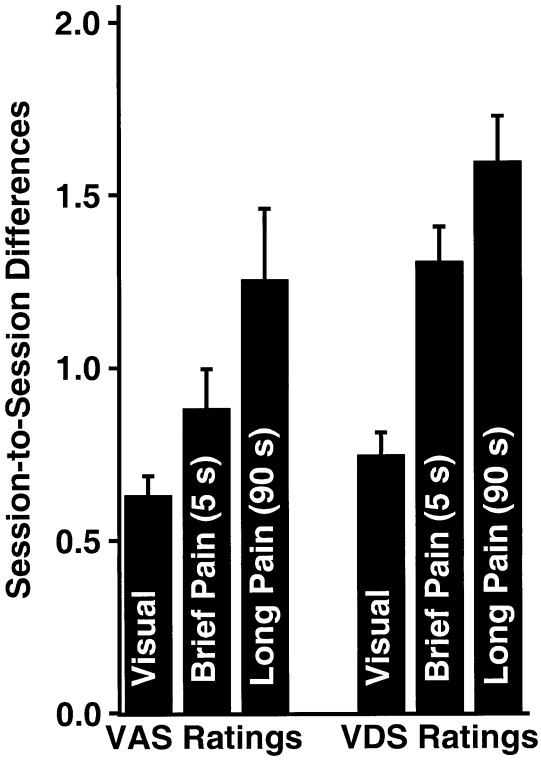

Fig. 7. Reproducibility of visual, brief,and prolonged heat stimuli based on intensity ratings. Ratings of the visual stimuli exhibited significantly smaller session-to-session differences than ratings of both the prolonged and brief heat pain stimuli.

the stimuli were delivered using a precise, well-calibrated, feedback-controlled device which was carefully positioned by an experienced investigator. Large session-to-session variations in heat pain ratings have been previously noted when similar, highly reproducible, feedback-controlled devices are employed (Yarnitsky et al., 1996). Similarly, stimulation of different skin sites between sessions is unlikely to account for a substantial portion of the temporal variability in the pain experience, since stimulus sites were held consistent between sessions to control for regional differences in skin sensitivity. Additionally, the sequential stimulation of multiple sites (i.e. prolonged heat stimulation), which would be predicted to minimize perceptual differences arising from regional differences in skin sensitivity, failed to reduce session-to-session variability of pain ratings obtained with either scale. Variations in external physical factors likely contributed only minimally to session-to-session variations in the pain experience. Although significant variations (±10°C) in ambient temperature have been shown to alter pain sensitivity, the possibility of such large fluctuations in ambient temperature were minimized by conducting testing in a thermostatically controlled, air-conditioned environment (Strigo et al., 2000).

More generalized physiological, psychological, and social variables unique to each individual may have contributed substantially to session-to-session variations in the actual experience of pain. Such variables have been shown to markedly alter the experience of pain (Price, 2000). In the case of physiological variables, small differences in baseline skin temperature evoked by differing activity levels prior to testing and/or variations in peripheral blood flow may have contributed to session-to-session variations in nociceptor activation (Wu et al., 2001). Furthermore, changing expectations about the experimental paradigm, fluctuating anxiety about the experimental stimuli, stress, and other distractors arising from changes in subject's daily life can all significantly alter the pain experience.

4.3. Temporal variations in pain scale usage

In addition to session-to-session variations in the actual pain experience, temporal variations in pain scale usage also contribute substantially to variations in pain ratings. If, as discussed above, a minimum of 65–72% of session-to-session variation in pain ratings is due to variations in the actual pain experience, then up to 28–35% of the remaining variability may be potentially attributed to variations in

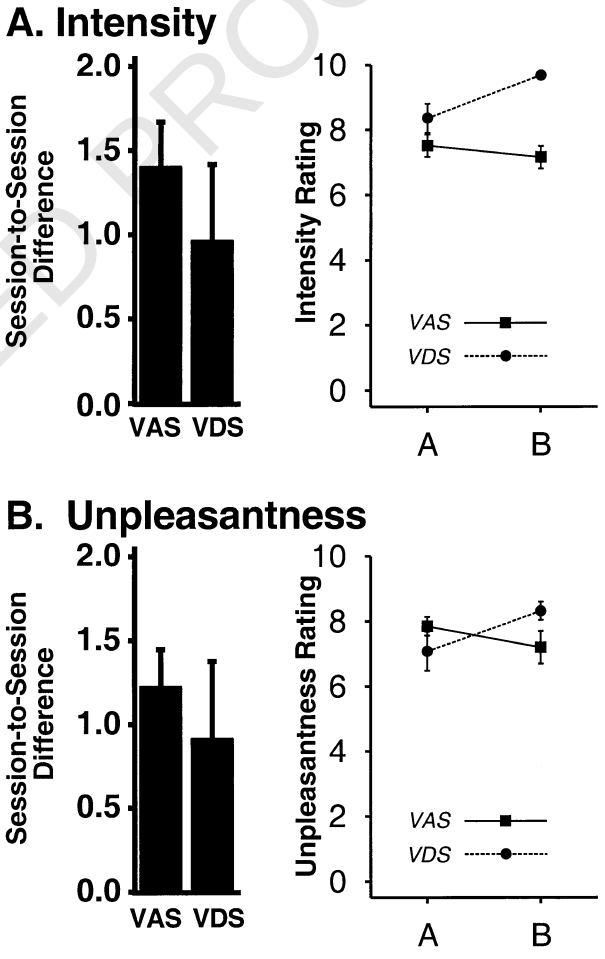

Fig. 8. Differences in worst pain ratings over time. Both VAS and VDS ratings remained unchanged over time and exhibited statistically indistinguishable session-to-session differences. ratings.

scale usage. Such variations would include a variety of response biases common to magnitude scaling procedures.

4.4. Verbal descriptor scales vs. visual analog scales

The central purpose of this investigation was to examine the reproducibility of the pain experience and measures of that experience, rather than to compare VAS and VDS methods. However, several aspects of these different scaling procedures deserve comment. First and most importantly, both methods were sensitive to small differences in noxious stimulus temperatures. The VDS performed surprisingly well in this regard, given that no spatial cues were provided to assist in the semantic interpretation of descriptor magnitude. Nevertheless, the VAS exhibited a slight advantage in sensitivity in that statistically significant differences were detected between all adjacent pairs of noxious stimulus intensities (1°C differences in the case of the short duration stimuli and 2°C differences in the case of the long duration stimuli). In contrast, somewhat fewer statistically significant differences were detected in pairwise analyses of VDS measures of 1°C differences in brief heat pain. Second, clear criterion shifts were evident with the VDS. Subjects significantly shifted their interpretation of the magnitude of the verbal descriptors from session-to-session, although in the case of pain intensity ratings, descriptor ordering tended to stabilize over time. It is important to note that more recent variants of VDS typically pair verbal descriptors with either numbers or VAS to minimize this potential response bias (Coghill and Gracely, 1996; Chibnall and Tait, 2001). Third, in the case of VDS ratings of pain affect, each subject used an average of only six of the 15 descriptors per session, with just four descriptors accounting for 78% of the responses. This apparent perseveration in affective descriptor choices likely contributed to the somewhat limited sensitivity of VDS to small differences in pain affect.

4.5. Recommendations to maximize reproducibility

Pain is an experience subject to substantial temporal variation. Thus, assessments of pain in both laboratory and clinical settings must be structured to extract as much useful information as possible despite this substantial temporal variation. The present findings suggest several methods for maximizing reproducibility:

1. In each subject, obtain multiple assessments of the same stimulus/state within a given session. Within-subject averaging of assessments substantially minimizes session-to-session variation. With both VAS and VDS ratings, session-to-session variation was reduced by approximately 27–33% when three assessments of brief heat pain were obtained instead of a single assessment.
2. Minimize the use of prolonged noxious stimuli. In cases where pain can be evoked, multiple presentations and assessments of brief stimuli produce more reproducible ratings than a single assessment of a single prolonged stimulus. Although easily accomplished in most experimental investigations of acute pain, the use of multiple, brief stimuli can also be employed in clinical/chronic pain states with some forms of evocable pain, allodynia, or hyperalgesia.
3. Provide subjects with a training period distinct from the study period. Order effects present with the VDS tended to decrease over time as subjects gained more experience with the scales.
4. Ensure that interpretation of scale parameters/descriptors remains constant over time. In general, repeated presentation of scale instructions may enhance stability of scale interpretation. Changes in the interpretation of scale magnitude can be rapidly evaluated by a relatively simple grey intensity rating task (as above, Fig. 1). In the case of category scales for pain magnitude, subjects' ordering of categories should be assessed during every testing session to ensure that re-interpretation does not occur.

Although implementation of these recommendations may help to minimize session-to-session variations in future investigations of pain, it is important to realize that a myriad of complex, temporally fluctuating variables in the lives of subjects and/or patients may exert profound effects on their pain experience. The development of an understanding of the neural mechanisms through which such psycho-physiological variables influence the pain experience will represent a crucial step in the understanding of how the conscious pain experience is constructed.

Acknowledgements

This research was funded by the NIDR Emerging Opportunities Fund. R.C.C. was supported by NINDS NS39426.

References

Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. The Statistician 1983;32:307–317.

Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–310.

British Standards Institution. Precision of test methods I: guide for the determination and reproducibility for a standard test method (BS 5497). London: BSI, 1979.

Chibnall JT, Tait RC. Pain assessment in cognitively impaired and unimpaired older adults: a comparison of four scales. Pain 2001;92:173–186.

Coghill RC, Gracely RH. Validation of the combined numerical/verbal descriptor scale for pain. Am Pain Soc Abstr 1996;15:A86.

Gracely RH, McGrath P, Dubner R. Ratio scales of sensory and affective verbal pain descriptors. Pain 1978;5:5–18.

Heft MW, Gracely RH, Dubner R, McGrath PA. A validation model for verbal description scaling of human clinical pain. Pain 1980;9:363–373.

Price DD. Comments on Yarnitsky et al., Pain, 67 (1996) 327–333. Pain 1997;73:108–109.

Price DD. Psychological and neural mechanisms of the affective dimension of pain. Science 2000;288:1769–1772.

Price DD, McGrath PA, Rafii A, Buckingham B. The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. Pain 1983;17:45–56.

Price DD, McHaffie GJ, Larson MA. Spatial summation of heat-induced pain: influence of stimulus area and spatial separation of stimuli on perceived pain sensation intensity and unpleasantness. J Neurophysiol 1989;62:1270–1279.

Price DD, Bush FM, Long S, Harkins SW. A comparison of pain measurement characteristics of mechanical visual analogue and simple numerical rating scales. Pain 1994;56:217–226.

Strigo IA, Carli F, Bushnell MC. Effect of ambient temperature on human pain and temperature perception. Anesthesiology 2000;92:699–707.

Wu G, Campbell JN, Meyer RA. Effects of baseline skin temperature on pain ratings to suprathreshold temperature-controlled stimuli. Pain 2001;90:151–156.

Yarnitsky D, Sprecher E, Zaslansky R, Hemli JA. Heat pain thresholds: normative data and repeatability. Pain 1995;60:329–332.

Yarnitsky D, Sprecher E, Zaslansky R, Hemli JA. Multiple session experimental pain measurement. Pain 1996;67:327–333.